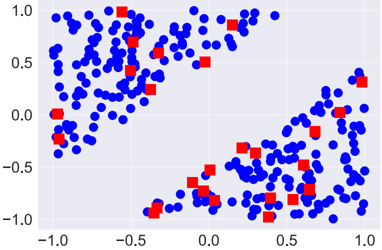
The hidden talents of logistic regression

From noisy labels to point processes

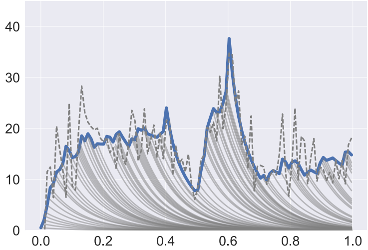
Aditya Krishna Menon

November 7th, 2017

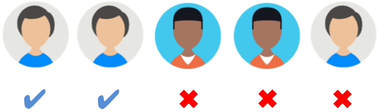
Three problems...



Label noise

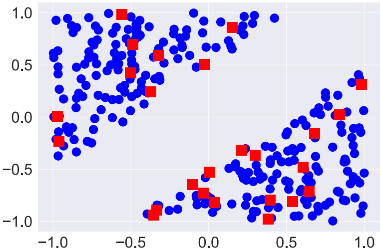


Point processes

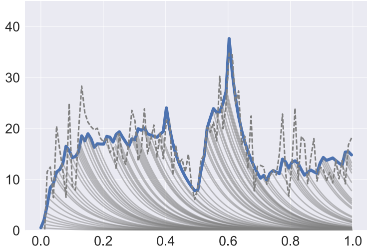


Fairness

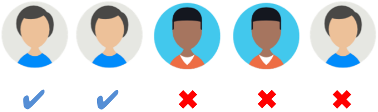
Three problems...one solution?



Label noise

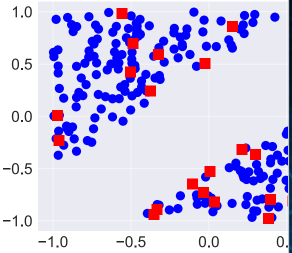


Point processes

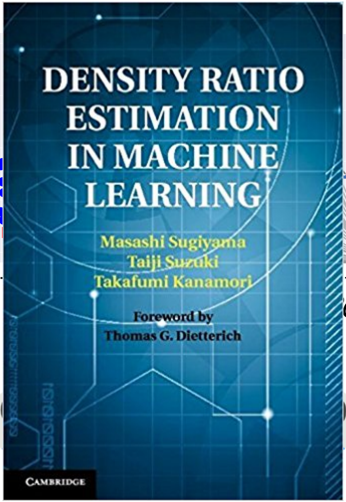


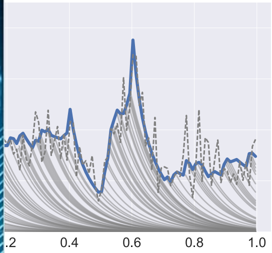
Fairness

Three problems...one solution?



Label noise

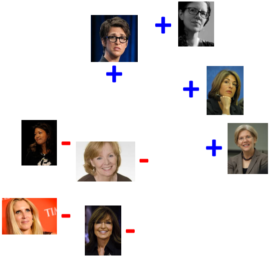




oint processes

Fairness

DRE applications

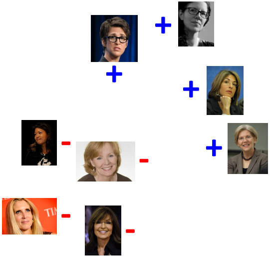


Covariate shift

Outlier detection

Robot transition estimation

DRE applications



Covariate shift

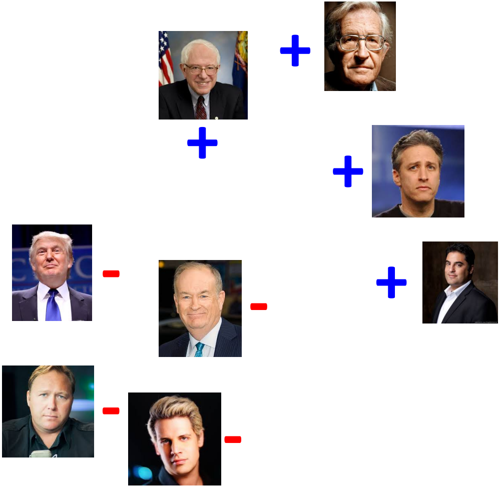
Outlier detection

Robot transition estimation

In some cases, a different view may be more natural

Class-probability estimation (CPE)

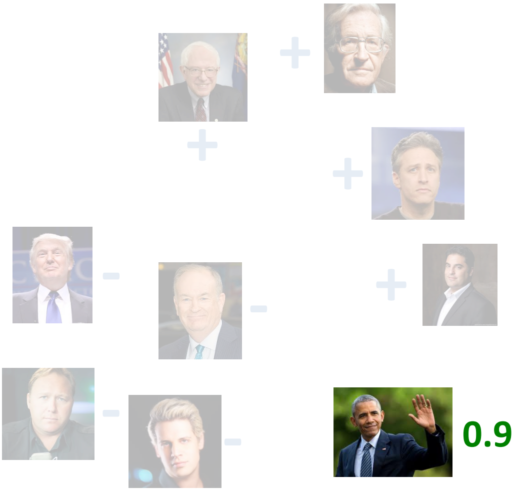
From labelled instances



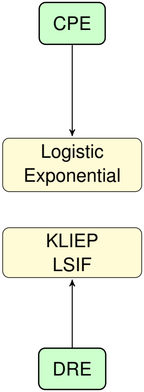
Class-probability estimation (CPE)

From labelled instances, estimate probability of instance being +'ve

- e.g. using logistic regression

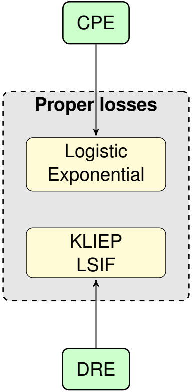


This talk



This talk

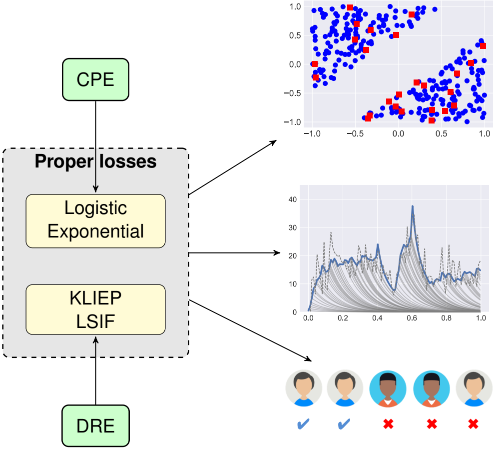
A formal link between DRE and CPE



This talk

A formal link between DRE and CPE

CPE approach to three distinct learning problems



Class-probability estimation

Distributions for learning with binary labels

Fix an instance space $\mathcal{X}$ (e.g. $\mathbb{R}^n$)

Distributions for learning with binary labels

Fix an instance space $\mathcal{X}$ (e.g. $\mathbb{R}^n$)

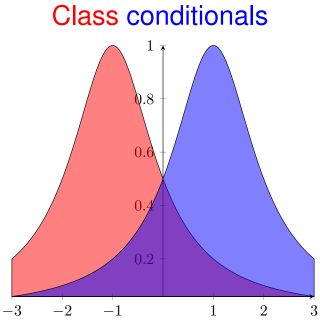
Let D be a distribution over $\mathcal{X} \times \{\pm 1\}$, with

Distributions for learning with binary labels

Fix an instance space $\mathcal{X}$ (e.g. $\mathbb{R}^n$)

Let D be a distribution over $\mathcal{X} \times \{\pm 1\}$, with

$$(P(x), Q(x)) = (\mathbb{P}(\mathsf{X} = x \mid \mathsf{Y} = +1), \mathbb{P}(\mathsf{X} = x \mid \mathsf{Y} = -1))$$



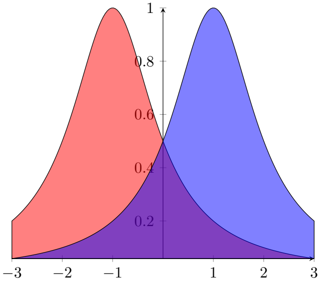
Distributions for learning with binary labels

Fix an instance space $\mathcal{X}$ (e.g. $\mathbb{R}^n$)

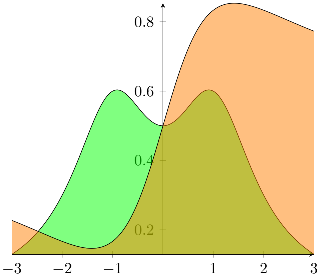
Let D be a distribution over $\mathcal{X} \times \{\pm 1\}$, with

$$(P(x), Q(x)) = (\mathbb{P}(\mathsf{X} = x \mid \mathsf{Y} = +1), \mathbb{P}(\mathsf{X} = x \mid \mathsf{Y} = -1))$$
$$(M(x), \eta(x)) = (\mathbb{P}(\mathsf{X} = x), \mathbb{P}(\mathsf{Y} = +1 \mid \mathsf{X} = x))$$

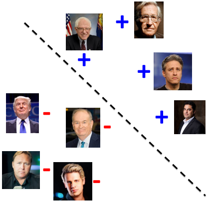
Class conditionals



Marginal and class-probability function



Scorers, losses, risks



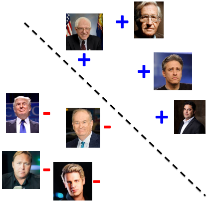
A scorer is any $s\colon \mathcal{X} \to \mathbb{R}$

- e.g. linear scorer $s\colon x \mapsto \langle w, x \rangle$

Scorers, losses, risks

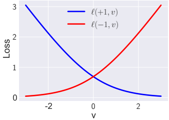
A scorer is any $s\colon \mathcal{X} \to \mathbb{R}$

- e.g. linear scorer $s\colon x \mapsto \langle w, x \rangle$



A loss is any $\ell\colon \{\pm 1\} \times \mathbb{R} \to \mathbb{R}_+$

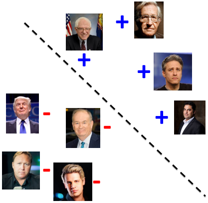
- e.g. logistic loss $\ell\colon (y, v) \mapsto \log(1 + e^{-yv})$



Scorers, losses, risks

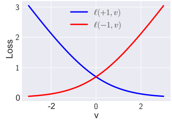
A scorer is any $s\colon \mathcal{X} \to \mathbb{R}$

- e.g. linear scorer $s\colon x \mapsto \langle w, x \rangle$



A loss is any $\ell\colon \{\pm 1\} \times \mathbb{R} \to \mathbb{R}_+$

- e.g. logistic loss $\ell\colon (y, v) \mapsto \log(1 + e^{-yv})$



The risk of scorer s wrt loss ℓ and distribution D is

$$\mathop{\mathbb{E}}_{(\mathsf{X},\mathsf{Y}) \sim D} \left[\ell(\mathsf{Y}, s(\mathsf{X})) \right]$$

- average loss on a random sample

Class-probability estimation

Goal: estimate $\eta(x) \doteq \mathbb{P}(\mathsf{Y} = +1 \mid \mathsf{X} = x)$

Class-probability estimation

Goal: estimate $\eta(x) \doteq \mathbb{P}(\mathsf{Y} = +1 \mid \mathsf{X} = x)$

For suitable $\mathcal{S} \subset \mathbb{R}^{\mathcal{X}}$, minimise empirical risk

$$\operatorname*{argmin}_{s \in \mathcal{S}} \frac{1}{N} \sum_{n=1}^{N} \ell(y_n, s(x_n))$$

for strictly proper composite ℓ

Class-probability estimation

Goal: estimate $\eta(x) \doteq \mathbb{P}(\mathsf{Y} = +1 \mid \mathsf{X} = x)$

For suitable $\mathcal{S} \subset \mathbb{R}^{\mathcal{X}}$, minimise empirical risk

$$\underset{s \in \mathcal{S}}{\operatorname{argmin}} \frac{1}{N} \sum_{n=1}^{N} \ell(y_n, s(x_n))$$

for strictly proper composite ℓ viz., for invertible link $\Psi : (0,1) \to \mathbb{R}$,

$$\underset{s \in \mathbb{R}^{\mathcal{X}}}{\operatorname{argmin}} \mathbb{E}\left[\ell(\mathsf{Y}, s(\mathsf{X}))\right] = \Psi \circ \eta$$

- e.g. for logistic loss, $\Psi(u) = \log \frac{u}{1-u}$

Class-probability estimation

Goal: estimate $\eta(x) \doteq \mathbb{P}(\mathsf{Y} = +1 \mid \mathsf{X} = x)$

For suitable $\mathcal{S} \subset \mathbb{R}^{\mathcal{X}}$, minimise empirical risk

$$\underset{s \in \mathcal{S}}{\operatorname{argmin}} \frac{1}{N} \sum_{n=1}^{N} \ell(y_n, s(x_n))$$

for strictly proper composite ℓ viz., for invertible link $\Psi : (0,1) \to \mathbb{R}$,

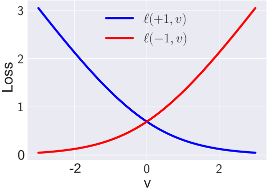
$$\underset{s \in \mathbb{R}^{\mathcal{X}}}{\operatorname{argmin}} \mathbb{E}\left[\ell(\mathsf{Y}, s(\mathsf{X}))\right] = \Psi \circ \eta$$

- e.g. for logistic loss, $\Psi(u) = \log \frac{u}{1-u}$

Estimate $\hat{\eta} \doteq \Psi^{-1} \circ s$

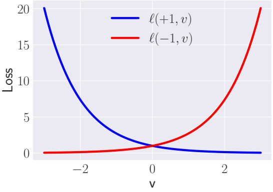
- e.g. for logistic loss, $\hat{\eta}(x) = 1/(1 + \exp(-s(x)))$

Examples of proper composite losses



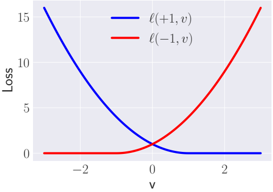
Logistic loss

$$\Psi^{-1}: v \mapsto 1/(1+\exp(-v))$$



Exponential loss

$$\Psi^{-1}: v \mapsto 1/(1+\exp(-2v))$$



Square hinge loss

$$\Psi^{-1}: v \mapsto \min(\max(0, (v+1)/2), 1)$$

Class-probabilities and density ratios

CPE versus DRE

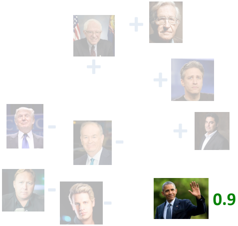
Given samples $\mathsf{S} \sim D^N$, with $D = (P, Q) = (M, \eta)$:

CPE versus DRE

Given samples S $\sim D^N$, with $D = (P, Q) = (M, \eta)$:

Class-probability estimation (CPE)
Estimate η

- class-probability function

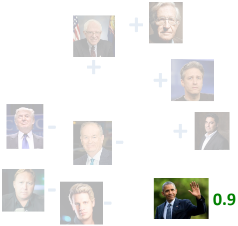


CPE versus DRE

Given samples S $\sim D^N$, with $D = (P, Q) = (M, \eta)$:

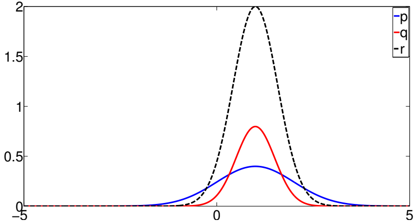
Class-probability estimation (CPE)
Estimate η

- class-probability function



Density ratio estimation (DRE)
Estimate $r = p/q$

- class-conditional density ratio



CPE and DRE: exact solutions

Bayes' rule shows DRE and CPE are linked (Bickel et al, 2009):

$$r(x) \doteq \frac{p(x)}{q(x)}$$

CPE and DRE: exact solutions

Bayes' rule shows DRE and CPE are linked (Bickel et al, 2009):

$$
\begin{aligned}
r(x) &\doteq \frac{p(x)}{q(x)} \\
&= \frac{\mathbb{P}(\mathsf{X} = x \mid \mathsf{Y} = +1)}{\mathbb{P}(\mathsf{X} = x \mid \mathsf{Y} = -1)}
\end{aligned}
$$

CPE and DRE: exact solutions

Bayes' rule shows DRE and CPE are linked (Bickel et al, 2009):

$$
\begin{aligned}
r(x) &\doteq \frac{p(x)}{q(x)} \\
&= \frac{\mathbb{P}(\mathsf{X} = x \mid \mathsf{Y} = +1)}{\mathbb{P}(\mathsf{X} = x \mid \mathsf{Y} = -1)} \\
&\propto \frac{\mathbb{P}(\mathsf{Y} = +1 \mid \mathsf{X} = x)}{\mathbb{P}(\mathsf{Y} = -1 \mid \mathsf{X} = x)}
\end{aligned}
$$

CPE and DRE: exact solutions

Bayes' rule shows DRE and CPE are linked (Bickel et al, 2009):

$$
\begin{aligned}
r(x) &\doteq \frac{p(x)}{q(x)} \\
&= \frac{\mathbb{P}(\mathsf{X} = x \mid \mathsf{Y} = +1)}{\mathbb{P}(\mathsf{X} = x \mid \mathsf{Y} = -1)} \\
&\propto \frac{\mathbb{P}(\mathsf{Y} = +1 \mid \mathsf{X} = x)}{\mathbb{P}(\mathsf{Y} = -1 \mid \mathsf{X} = x)} \\
&= \frac{\eta(x)}{1 - \eta(x)}
\end{aligned}
$$

CPE and DRE: exact solutions

Bayes' rule shows DRE and CPE are linked (Bickel et al, 2009):

$$
\begin{aligned}
r(x) &\doteq \frac{p(x)}{q(x)} \\
&= \frac{\mathbb{P}(\mathsf{X} = x \mid \mathsf{Y} = +1)}{\mathbb{P}(\mathsf{X} = x \mid \mathsf{Y} = -1)} \\
&\propto \frac{\mathbb{P}(\mathsf{Y} = +1 \mid \mathsf{X} = x)}{\mathbb{P}(\mathsf{Y} = -1 \mid \mathsf{X} = x)} \\
&= \frac{\eta(x)}{1 - \eta(x)}
\end{aligned}
$$

Obtain η via CPE $\rightarrow$ also obtain r for DRE

CPE and DRE: exact solutions

Bayes' rule shows DRE and CPE are linked (Bickel et al, 2009):

$$
\begin{aligned}
r(x) &\doteq \frac{p(x)}{q(x)} \\
&= \frac{\mathbb{P}(\mathsf{X} = x \mid \mathsf{Y} = +1)}{\mathbb{P}(\mathsf{X} = x \mid \mathsf{Y} = -1)} \\
&\propto \frac{\mathbb{P}(\mathsf{Y} = +1 \mid \mathsf{X} = x)}{\mathbb{P}(\mathsf{Y} = -1 \mid \mathsf{X} = x)} \\
&= \frac{\eta(x)}{1 - \eta(x)}
\end{aligned}
$$

Obtain η via CPE $\rightarrow$ also obtain r for DRE

But what about approximate solutions?

CPE and DRE: approximate solutions?

Natural class-probability estimate: $\hat{\eta} \doteq \Psi^{-1} \circ s$

- e.g. for logistic loss, $\hat{\eta}(x) = 1/(1 + e^{-s(x)})$

CPE and DRE: approximate solutions?

Natural class-probability estimate: $\hat{\eta} \doteq \Psi^{-1} \circ s$

- e.g. for logistic loss, $\hat{\eta}(x) = 1/(1 + e^{-s(x)})$

Natural density ratio estimate:

$$\hat{r}(x) \doteq \frac{\hat{\eta}(x)}{1 - \hat{\eta}(x)}$$

- e.g. for logistic loss, $\hat{r}(x) = e^{s(x)}$

CPE and DRE: approximate solutions?

Natural class-probability estimate: $\hat{\eta} \doteq \Psi^{-1} \circ s$

- e.g. for logistic loss, $\hat{\eta}(x) = 1/(1 + e^{-s(x)})$

Natural density ratio estimate:

$$\hat{r}(x) \doteq \frac{\hat{\eta}(x)}{1 - \hat{\eta}(x)}$$

- e.g. for logistic loss, $\hat{r}(x) = e^{s(x)}$

Intuitive, but what can we guarantee about this?

CPE as Bregman minimisation

For proper composite ℓ, the regret of a scorer is

$$\mathrm{reg}(s) \doteq \mathbb{E}\left[\ell(\mathsf{Y}, s(\mathsf{X}))\right] - \min_{\bar{s} \in \mathbb{R}^{\mathcal{X}}} \mathbb{E}\left[\ell(\mathsf{Y}, \bar{s}(\mathsf{X}))\right]$$

CPE as Bregman minimisation

For proper composite ℓ, the regret of a scorer is

$$\begin{aligned}\operatorname{reg}(s) &\doteq \mathbb{E}\left[\ell(\mathsf{Y}, s(\mathsf{X}))\right] - \min_{\bar{s} \in \mathbb{R}^{\mathcal{X}}} \mathbb{E}\left[\ell(\mathsf{Y}, \bar{s}(\mathsf{X}))\right] \\ &= \mathbb{E}\left[B_f(\eta(\mathsf{X}), \hat{\eta}(\mathsf{X}))\right]\end{aligned}$$

for Bregman divergence $B_f(\cdot, \cdot)$ and loss-specific f

CPE as Bregman minimisation

For proper composite ℓ, the regret of a scorer is

$$\begin{aligned} \mathrm{reg}(s) &\doteq \mathbb{E}\left[\ell(\mathsf{Y}, s(\mathsf{X}))\right] - \min_{\bar{s} \in \mathbb{R}^{\mathcal{X}}} \mathbb{E}\left[\ell(\mathsf{Y}, \bar{s}(\mathsf{X}))\right] \\ &= \mathbb{E}\left[B_f(\eta(\mathsf{X}), \hat{\eta}(\mathsf{X}))\right] \end{aligned}$$

for Bregman divergence $B_f(\cdot, \cdot)$ and loss-specific f

- e.g. for logistic loss, regret is a KL projection

$$\mathrm{reg}(s) = \mathbb{E}\left[\mathrm{KL}(\eta(\mathsf{X}) \| \hat{\eta}(\mathsf{X}))\right]$$

CPE as Bregman minimisation

For proper composite ℓ, the regret of a scorer is

$$\begin{aligned} \mathrm{reg}(s) &\doteq \mathbb{E}\left[\ell(\mathsf{Y}, s(\mathsf{X}))\right] - \min_{\bar{s} \in \mathbb{R}^{\mathcal{X}}} \mathbb{E}\left[\ell(\mathsf{Y}, \bar{s}(\mathsf{X}))\right] \\ &= \mathbb{E}\left[B_f(\eta(\mathsf{X}), \hat{\eta}(\mathsf{X}))\right] \end{aligned}$$

for Bregman divergence $B_f(\cdot, \cdot)$ and loss-specific f

- e.g. for logistic loss, regret is a KL projection

$$\mathrm{reg}(s) = \mathbb{E}\left[\mathsf{KL}(\eta(\mathsf{X}) \| \hat{\eta}(\mathsf{X}))\right]$$

Concrete sense in which estimate $\hat{\eta}$ is reasonable

CPE as Bregman minimisation

For proper composite ℓ, the regret of a scorer is

$$\begin{aligned}\operatorname{reg}(s) &\doteq \mathbb{E}\left[\ell(\mathsf{Y}, s(\mathsf{X}))\right] - \min_{\bar{s} \in \mathbb{R}^{\mathcal{X}}} \mathbb{E}\left[\ell(\mathsf{Y}, \bar{s}(\mathsf{X}))\right] \\ &= \mathbb{E}\left[B_f(\eta(\mathsf{X}), \hat{\eta}(\mathsf{X}))\right]\end{aligned}$$

for Bregman divergence $B_f(\cdot, \cdot)$ and loss-specific f

- e.g. for logistic loss, regret is a KL projection

$$\operatorname{reg}(s) = \mathbb{E}\left[\mathsf{KL}(\eta(\mathsf{X}) \| \hat{\eta}(\mathsf{X}))\right]$$

Concrete sense in which estimate $\hat{\eta}$ is reasonable

Is there a similar sense in which $\hat{r}$ is reasonable?

A Bregman identity

One can show that:

$$(\forall x, y \in [0,1])\, B_f\,(x, y) =$$

A Bregman identity

One can show that:

$$(\forall x, y \in [0,1])\, B_f(x,y) = (1-x) \cdot B_{f^\circledast}\left(\frac{x}{1-x}, \frac{y}{1-y}\right),$$

where $f^\circledast : z \mapsto (1+z) \cdot f\left(\frac{z}{1+z}\right)$

Menon and Ong, Linking losses for density ratio and class-probability estimation, ICML 2016.

A Bregman identity

One can show that:

$$(\forall x, y \in [0,1])\, B_f(x, y) = (1 - x) \cdot B_{f^{\circledast}}\left(\frac{x}{1-x}, \frac{y}{1-y}\right),$$

where $f^{\circledast} : z \mapsto (1+z) \cdot f\left(\frac{z}{1+z}\right)$

Since $r = \frac{\eta}{1-\eta}$ and $\hat{r} = \frac{\hat{\eta}}{1-\hat{\eta}}$,

$$\mathrm{reg}(s) = \mathop{\mathbb{E}}_{\mathsf{X} \sim M}\left[B_f(\eta(\mathsf{X}), \hat{\eta}(\mathsf{X}))\right]$$

Menon and Ong, Linking losses for density ratio and class-probability estimation, ICML 2016.

A Bregman identity

One can show that:

$$(\forall x, y \in [0,1]) \, B_f(x,y) = (1-x) \cdot B_{f^{\circledast}}\left(\frac{x}{1-x}, \frac{y}{1-y}\right),$$

where $f^{\circledast} : z \mapsto (1+z) \cdot f\left(\frac{z}{1+z}\right)$

Since $r = \frac{\eta}{1-\eta}$ and $\hat{r} = \frac{\hat{\eta}}{1-\hat{\eta}}$,

$$\mathrm{reg}(s) = \mathop{\mathbb{E}}_{\mathsf{X} \sim M}\left[B_f(\eta(\mathsf{X}), \hat{\eta}(\mathsf{X}))\right] \propto \mathop{\mathbb{E}}_{\mathsf{X} \sim Q}\left[B_{f^{\circledast}}(r(\mathsf{X}), \hat{r}(\mathsf{X}))\right]$$

Menon and Ong, Linking losses for density ratio and class-probability estimation, ICML 2016.

A Bregman identity

One can show that:

$$(\forall x, y \in [0,1])\, B_f(x,y) = (1-x) \cdot B_{f^{\circledast}}\left(\frac{x}{1-x}, \frac{y}{1-y}\right),$$

where $f^{\circledast} : z \mapsto (1+z) \cdot f\left(\frac{z}{1+z}\right)$

Since $r = \frac{\eta}{1-\eta}$ and $\hat{r} = \frac{\hat{\eta}}{1-\hat{\eta}}$,

$$\mathrm{reg}(s) = \mathop{\mathbb{E}}_{\mathsf{X} \sim M}\left[B_f(\eta(\mathsf{X}), \hat{\eta}(\mathsf{X}))\right] \propto \mathop{\mathbb{E}}_{\mathsf{X} \sim Q}\left[B_{f^{\circledast}}(r(\mathsf{X}), \hat{r}(\mathsf{X}))\right]$$

CPE implicitly estimates density ratios

- complementary to (Sugiyama et al., 2012)

Menon and Ong, Linking losses for density ratio and class-probability estimation, ICML 2016.

CPE and DRE: summary

The asymptotic targets of CPE and DRE are closely linked

- $r = \frac{\eta}{1-\eta}$

Nock et al. A scaled Bregman theorem with applications. NIPS 2016.

CPE and DRE: summary

The asymptotic targets of CPE and DRE are closely linked

- $r = \frac{\eta}{1-\eta}$

Estimating $\hat{\eta}$ by proper loss minimisation is reasonable

- minimises Bregman divergence to η

Nock et al. A scaled Bregman theorem with applications. NIPS 2016.

CPE and DRE: summary

The asymptotic targets of CPE and DRE are closely linked

- $r = \frac{\eta}{1-\eta}$

Estimating $\hat{\eta}$ by proper loss minimisation is reasonable

- minimises Bregman divergence to η

Estimating $\hat{r}$ by proper loss minimisation is reasonable

- minimises alternate Bregman divergence to r

Nock et al. A scaled Bregman theorem with applications. NIPS 2016.

CPE and DRE: summary

The asymptotic targets of CPE and DRE are closely linked

- $r = \frac{\eta}{1-\eta}$

Estimating $\hat{\eta}$ by proper loss minimisation is reasonable

- minimises Bregman divergence to η

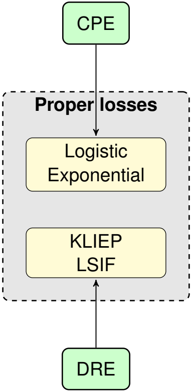
Estimating $\hat{r}$ by proper loss minimisation is reasonable

- minimises alternate Bregman divergence to r

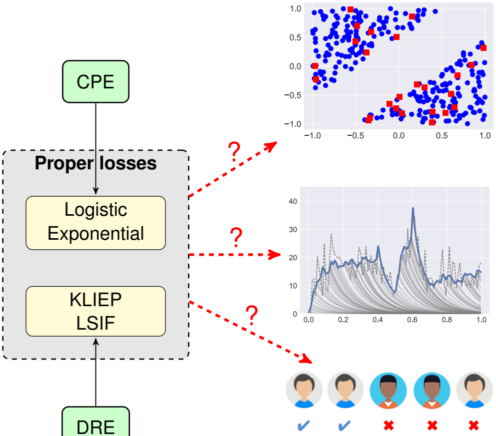
Underlying Bregman identity has multi-dimensional generalisation

Nock et al. A scaled Bregman theorem with applications. NIPS 2016.

Summary thus far

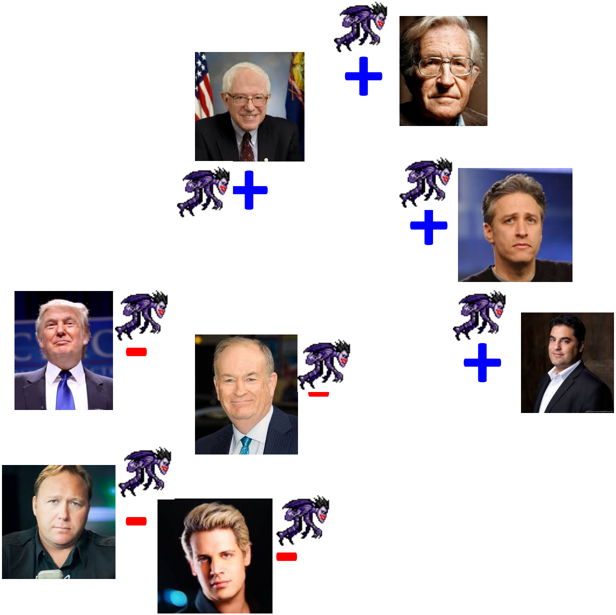


Summary thus far

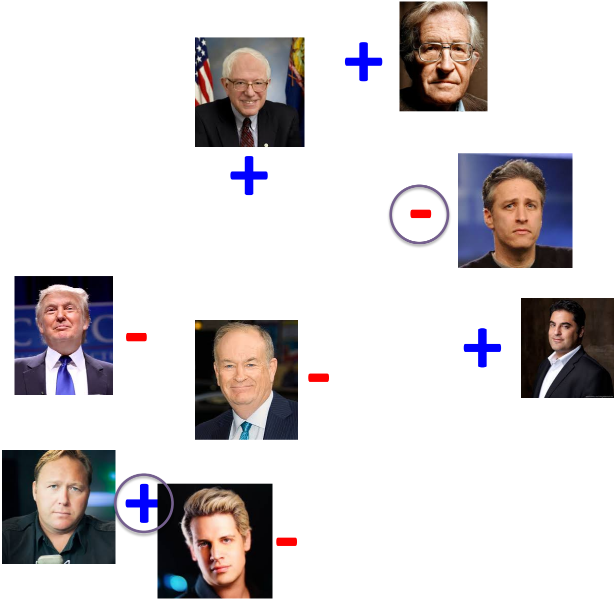


Learning from noisy binary labels

Learning from noisy binary labels



Learning from noisy binary labels



Label noise: formally

We care about "clean" D

Ideal

$$\min_s \mathop{\mathbb{E}}_{(\mathsf{X},\mathsf{Y})} [\ell(\mathsf{Y}, s(\mathsf{X}))]$$

Label noise: formally

We care about "clean" D, but observe samples from $\bar{D} \neq D$

- $\mathbb{P}(\mathsf{X} = x)$ remains same, but $\mathbb{P}(\mathsf{Y} = +1 \mid \mathsf{X} = x)$ does not

Ideal

$$\min_s \mathop{\mathbb{E}}_{(\mathsf{X},\mathsf{Y})} \left[\ell(\mathsf{Y}, s(\mathsf{X}))\right]$$

Reality

$$\min_s \mathop{\mathbb{E}}_{(\mathsf{X},\bar{\mathsf{Y}})} \left[\ell(\bar{\mathsf{Y}}, s(\mathsf{X}))\right]$$

Label noise: formally

We care about "clean" D, but observe samples from $\bar{D} \neq D$

- $\mathbb{P}(\mathsf{X} = x)$ remains same, but $\mathbb{P}(\mathsf{Y} = +1 \mid \mathsf{X} = x)$ does not

Ideal

$$\min_s \mathop{\mathbb{E}}_{(\mathsf{X},\mathsf{Y})} [\ell(\mathsf{Y}, s(\mathsf{X}))]$$

Reality

$$\min_s \mathop{\mathbb{E}}_{(\mathsf{X},\bar{\mathsf{Y}})} \left[\ell(\bar{\mathsf{Y}}, s(\mathsf{X}))\right]$$

How to minimise the ideal risk?

Label noise: a CPE perspective

Denote by $\eta(x)$ the "clean" $\mathbb{P}(\mathsf{Y} = +1 \mid \mathsf{X} = x)$

Label noise: a CPE perspective

Denote by $\eta(x)$ the "clean" $\mathbb{P}(\mathsf{Y} = +1 \mid \mathsf{X} = x)$

Suppose (x, y) has label flipped with probability $\rho \in [0, 1/2)$

Label noise: a CPE perspective

Denote by $\eta(x)$ the "clean" $\mathbb{P}(\mathsf{Y} = +1 \mid \mathsf{X} = x)$

Suppose (x, y) has label flipped with probability $\rho \in [0, 1/2)$

The "noisy" $\mathbb{P}(\overline{\mathsf{Y}} = +1 \mid \mathsf{X} = x)$ is

$$\bar{\eta}(x) = (1 - \rho) \cdot \eta(x) + \rho \cdot (1 - \eta(x))$$

Label noise: a CPE perspective

Denote by $\eta(x)$ the "clean" $\mathbb{P}(\mathsf{Y} = +1 \mid \mathsf{X} = x)$

Suppose (x, y) has label flipped with probability $\rho \in [0, 1/2)$

The "noisy" $\mathbb{P}(\bar{\mathsf{Y}} = +1 \mid \mathsf{X} = x)$ is

$$\bar{\eta}(x) = (1 - \rho) \cdot \eta(x) + \rho \cdot (1 - \eta(x))$$

We may write

$$\begin{bmatrix} \bar{\eta}(x) \\ 1 - \bar{\eta}(x) \end{bmatrix} = \underbrace{\begin{bmatrix} 1 - \rho & \rho \\ \rho & 1 - \rho \end{bmatrix}}_{T} \begin{bmatrix} \eta(x) \\ 1 - \eta(x) \end{bmatrix}$$

Noise-corrected losses

$$\mathop{\mathbb{E}}_{(\mathsf{X},\mathsf{Y})} \left[\ell(\mathsf{Y}, s(\mathsf{X})) \right]$$

Noise-corrected losses

$$\mathop{\mathbb{E}}_{(\mathsf{X},\mathsf{Y})}\left[\ell(\mathsf{Y}, s(\mathsf{X}))\right] = \mathop{\mathbb{E}}_{\mathsf{X}\sim M}\begin{bmatrix}\eta(\mathsf{X}) \\ 1-\eta(\mathsf{X})\end{bmatrix}^{\mathrm{T}}\begin{bmatrix}\ell(+1, s(\mathsf{X})) & \ell(-1, s(\mathsf{X}))\end{bmatrix}$$

Noise-corrected losses

$$\mathop{\mathbb{E}}_{(\mathsf{X},\mathsf{Y})}[\ell(\mathsf{Y},s(\mathsf{X}))] = \mathop{\mathbb{E}}_{\mathsf{X}\sim M}\begin{bmatrix}\eta(\mathsf{X}) \\ 1-\eta(\mathsf{X})\end{bmatrix}^{\mathrm{T}}\begin{bmatrix}\ell(+1,s(\mathsf{X})) & \ell(-1,s(\mathsf{X}))\end{bmatrix}$$

$$= \mathop{\mathbb{E}}_{\mathsf{X}\sim M}\begin{bmatrix}\bar{\eta}(\mathsf{X}) \\ 1-\bar{\eta}(\mathsf{X})\end{bmatrix}^{\mathrm{T}}(T^{-1})^{\mathrm{T}}\begin{bmatrix}\ell(+1,s(\mathsf{X})) & \ell(-1,s(\mathsf{X}))\end{bmatrix}$$

Noise-corrected losses

$$\begin{aligned}
\mathop{\mathbb{E}}_{(\mathsf{X},\mathsf{Y})}\left[\ell(\mathsf{Y}, s(\mathsf{X}))\right] &= \mathop{\mathbb{E}}_{\mathsf{X}\sim M}\begin{bmatrix}\eta(\mathsf{X}) \\ 1-\eta(\mathsf{X})\end{bmatrix}^{\mathrm{T}}\begin{bmatrix}\ell(+1, s(\mathsf{X})) & \ell(-1, s(\mathsf{X}))\end{bmatrix} \\
&= \mathop{\mathbb{E}}_{\mathsf{X}\sim M}\begin{bmatrix}\bar{\eta}(\mathsf{X}) \\ 1-\bar{\eta}(\mathsf{X})\end{bmatrix}^{\mathrm{T}}(T^{-1})^{\mathrm{T}}\begin{bmatrix}\ell(+1, s(\mathsf{X})) & \ell(-1, s(\mathsf{X}))\end{bmatrix} \\
&= \mathop{\mathbb{E}}_{\mathsf{X}\sim M}\begin{bmatrix}\bar{\eta}(\mathsf{X}) \\ 1-\bar{\eta}(\mathsf{X})\end{bmatrix}^{\mathrm{T}}\begin{bmatrix}\bar{\ell}(+1, s(\mathsf{X})) & \bar{\ell}(-1, s(\mathsf{X}))\end{bmatrix}
\end{aligned}$$

for the noise-corrected loss (Natarajan et al., 2013)

$$\bar{\ell}(y, v) = \frac{1}{1-2\cdot\rho}\left((1-\rho)\cdot\ell(y, v) - \rho\cdot\ell(-y, v)\right)$$

Noise-corrected losses

$$
\begin{aligned}
\mathop{\mathbb{E}}_{(\mathsf{X},\mathsf{Y})}\left[\ell(\mathsf{Y}, s(\mathsf{X}))\right] &= \mathop{\mathbb{E}}_{\mathsf{X}\sim M}\begin{bmatrix}\eta(\mathsf{X}) \\ 1-\eta(\mathsf{X})\end{bmatrix}^{\mathrm{T}}\begin{bmatrix}\ell(+1, s(\mathsf{X})) & \ell(-1, s(\mathsf{X}))\end{bmatrix} \\
&= \mathop{\mathbb{E}}_{\mathsf{X}\sim M}\begin{bmatrix}\bar{\eta}(\mathsf{X}) \\ 1-\bar{\eta}(\mathsf{X})\end{bmatrix}^{\mathrm{T}}(T^{-1})^{\mathrm{T}}\begin{bmatrix}\ell(+1, s(\mathsf{X})) & \ell(-1, s(\mathsf{X}))\end{bmatrix} \\
&= \mathop{\mathbb{E}}_{\mathsf{X}\sim M}\begin{bmatrix}\bar{\eta}(\mathsf{X}) \\ 1-\bar{\eta}(\mathsf{X})\end{bmatrix}^{\mathrm{T}}\begin{bmatrix}\bar{\ell}(+1, s(\mathsf{X})) & \bar{\ell}(-1, s(\mathsf{X}))\end{bmatrix} \\
&= \mathop{\mathbb{E}}_{(\mathsf{X},\bar{\mathsf{Y}})}\left[\bar{\ell}(\bar{\mathsf{Y}}, s(\mathsf{X}))\right],
\end{aligned}
$$

for the noise-corrected loss (Natarajan et al., 2013)

$$
\bar{\ell}(y, v) = \frac{1}{1 - 2\cdot\rho}\left((1-\rho)\cdot\ell(y, v) - \rho\cdot\ell(-y, v)\right)
$$

Noise-corrected losses

$$
\begin{aligned}
\mathop{\mathbb{E}}_{(\mathsf{X},\mathsf{Y})}[\ell(\mathsf{Y}, s(\mathsf{X}))] &= \mathop{\mathbb{E}}_{\mathsf{X}\sim M}\begin{bmatrix}\eta(\mathsf{X}) \\ 1-\eta(\mathsf{X})\end{bmatrix}^{\mathrm{T}}\begin{bmatrix}\ell(+1, s(\mathsf{X})) & \ell(-1, s(\mathsf{X}))\end{bmatrix} \\
&= \mathop{\mathbb{E}}_{\mathsf{X}\sim M}\begin{bmatrix}\bar{\eta}(\mathsf{X}) \\ 1-\bar{\eta}(\mathsf{X})\end{bmatrix}^{\mathrm{T}}(T^{-1})^{\mathrm{T}}\begin{bmatrix}\ell(+1, s(\mathsf{X})) & \ell(-1, s(\mathsf{X}))\end{bmatrix} \\
&= \mathop{\mathbb{E}}_{\mathsf{X}\sim M}\begin{bmatrix}\bar{\eta}(\mathsf{X}) \\ 1-\bar{\eta}(\mathsf{X})\end{bmatrix}^{\mathrm{T}}\begin{bmatrix}\bar{\ell}(+1, s(\mathsf{X})) & \bar{\ell}(-1, s(\mathsf{X}))\end{bmatrix} \\
&= \mathop{\mathbb{E}}_{(\mathsf{X},\bar{\mathsf{Y}})}\left[\bar{\ell}(\bar{\mathsf{Y}}, s(\mathsf{X}))\right],
\end{aligned}
$$

for the noise-corrected loss (Natarajan et al., 2013)

$$
\bar{\ell}(y, v) = \frac{1}{1-2\cdot\rho}\left((1-\rho)\cdot\ell(y, v) - \rho\cdot\ell(-y, v)\right)
$$

But ρ is unknown...

Noise rate estimation

One can avoid knowing ρ for suitable ℓ

- eigenfunctions for the loss transform, e.g. “un-hinged” loss

van Rooyen et al. Learning with symmetric label noise: the importance of being unhinged. NIPS 2015.
Menon et al. Learning from corrupted binary labels via class-probability estimation. ICML 2015.

Noise rate estimation

One can avoid knowing ρ for suitable ℓ

- eigenfunctions for the loss transform, e.g. "un-hinged" loss

Alternately, assume $\min_x \eta(x) = 0$, $\max_x \eta(x) = 1$

- "guaranteed" positive and negative instances
- c.f. (Scott et al., 2013), (du Plessis et al., 2014)

van Rooyen et al. Learning with symmetric label noise: the importance of being unhinged. NIPS 2015.
Menon et al. Learning from corrupted binary labels via class-probability estimation. ICML 2015.

Noise rate estimation

One can avoid knowing ρ for suitable ℓ

- eigenfunctions for the loss transform, e.g. "un-hinged" loss

Alternately, assume $\min_x \eta(x) = 0$, $\max_x \eta(x) = 1$

- "guaranteed" positive and negative instances
- c.f. (Scott et al., 2013), (du Plessis et al., 2014)

Since $\bar{\eta}(x) = (1 - 2 \cdot \rho) \cdot \eta(x) + \rho$,

$$\min_x \bar{\eta}(x) = \rho \qquad \max_x \bar{\eta}(x) = 1 - \rho$$

van Rooyen et al. Learning with symmetric label noise: the importance of being unhinged. NIPS 2015.
Menon et al. Learning from corrupted binary labels via class-probability estimation. ICML 2015.

Noise rate estimation

One can avoid knowing ρ for suitable ℓ

- eigenfunctions for the loss transform, e.g. "un-hinged" loss

Alternately, assume $\min_x \eta(x) = 0$, $\max_x \eta(x) = 1$

- "guaranteed" positive and negative instances
- c.f. (Scott et al., 2013), (du Plessis et al., 2014)

Since $\bar{\eta}(x) = (1 - 2 \cdot \rho) \cdot \eta(x) + \rho$,

$$\min_x \bar{\eta}(x) = \rho \qquad \max_x \bar{\eta}(x) = 1 - \rho$$

Range of $\bar{\eta}$ lets us estimate ρ!

van Rooyen et al. Learning with symmetric label noise: the importance of being unhinged. NIPS 2015.
Menon et al. Learning from corrupted binary labels via class-probability estimation. ICML 2015.

Beyond symmetric binary noise

For asymmetric multi-class noise, we similarly have

$$\bar{\eta}(x) = T\eta(x)$$

where e.g. $\bar{\eta}(x) = (\mathbb{P}(\mathsf{Y} = 1 \mid \mathsf{X} = x), \ldots, \mathbb{P}(\mathsf{Y} = K \mid \mathsf{X} = x))$

- analogous noise-corrected loss and noise estimation

Patrini et al. Making deep neural networks robust to label noise: a loss correction approach. CVPR 2017.

Beyond symmetric binary noise

For asymmetric multi-class noise, we similarly have

$$\bar{\eta}(x) = T\eta(x)$$

where e.g. $\bar{\eta}(x) = (\mathbb{P}(\mathsf{Y} = 1 \mid \mathsf{X} = x), \ldots, \mathbb{P}(\mathsf{Y} = K \mid \mathsf{X} = x))$

- analogous noise-corrected loss and noise estimation

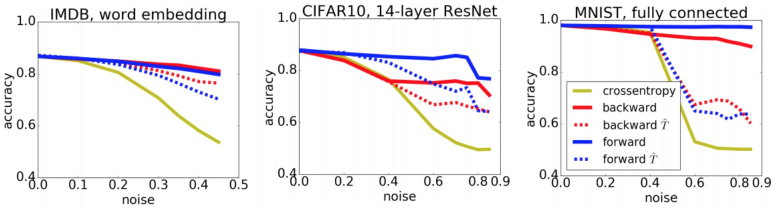
Broader range of weakly supervised problems captured

- confer (van Rooyen & Williamson, 2017)

Patrini et al. Making deep neural networks robust to label noise: a loss correction approach. CVPR 2017.

Illustration: deep network

Corrected losses with and without noise estimation



Patrini et al. Making deep neural networks robust to label noise: a loss correction approach. CVPR 2017.

Instance-dependent noise?

Denote by $\eta(x)$ the "clean" $\mathbb{P}(\mathsf{Y} = +1 \mid \mathsf{X} = x)$

Instance-dependent noise?

Denote by $\eta(x)$ the "clean" $\mathbb{P}(\mathsf{Y} = +1 \mid \mathsf{X} = x)$

Suppose (x, y) has label flipped with probability $\rho(x) \in [0, 1/2)$

Instance-dependent noise?

Denote by $\eta(x)$ the "clean" $\mathbb{P}(\mathsf{Y} = +1 \mid \mathsf{X} = x)$

Suppose (x, y) has label flipped with probability $\rho(x) \in [0, 1/2)$

The "noisy" class-probability function is

$$\bar{\eta}(x) = (1 - \rho(x)) \cdot \eta(x) + \rho(x) \cdot (1 - \eta(x))$$

Instance-dependent noise?

Denote by $\eta(x)$ the "clean" $\mathbb{P}(\mathsf{Y} = +1 \mid \mathsf{X} = x)$

Suppose (x, y) has label flipped with probability $\rho(x) \in [0, 1/2)$

The "noisy" class-probability function is

$$\bar{\eta}(x) = (1 - \rho(x)) \cdot \eta(x) + \rho(x) \cdot (1 - \eta(x))$$

Estimating $\rho(x)$ is non-trivial

- To make progress, we impose some structure on ρ and η

Assumptions on noise and distribution

Noise increases as $\eta(x)$ approaches $1/2$

- higher inherent uncertainty $\rightarrow$ higher noise

Assumptions on noise and distribution

Noise increases as $\eta(x)$ approaches $1/2$

- higher inherent uncertainty $\rightarrow$ higher noise

Class-probability is expressible as

$$\eta(x) = u(\langle w^*, x \rangle)$$

for some non-decreasing, Lipschitz $u(\cdot)$

- u unknown $\rightarrow$ single index model (SIM)
- such models learnable via Isotron (Kalai & Sastry, 2009)

Structure of noisy class-probability

Under these assumptions, one may show

$$\bar{\eta}(x) = \bar{u}(\langle w^*, x \rangle)$$

for monotone $\bar{u}$

- still in the SIM family!
- noise is baked into $\bar{u}$

Menon et al. Learning from binary labels with instance-dependent corruption. https://arxiv.org/abs/1605.00751

Structure of noisy class-probability

Under these assumptions, one may show

$$\bar{\eta}(x) = \bar{u}(\langle w^*, x \rangle)$$

for monotone $\bar{u}$

- still in the SIM family!
- noise is baked into $\bar{u}$

One can estimate $\bar{\eta}$ via Isotron

- do not need to know flip function ρ or link function u

Menon et al. Learning from binary labels with instance-dependent corruption. https://arxiv.org/abs/1605.00751

Illustration: instance-dependent noise

Label flip function $f(z) = (1 + e^{|z|/\alpha})^{-1}$

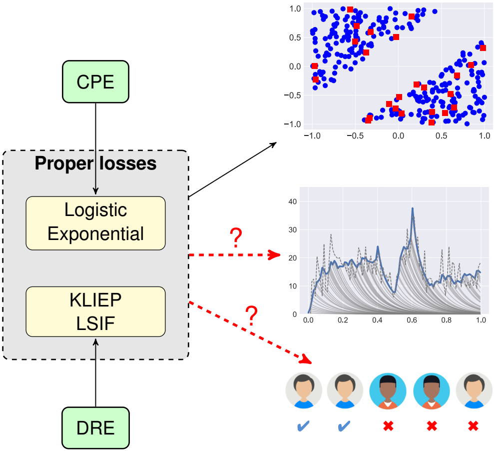
α	Ridge ACC	Isotron ACC
$\frac{1}{8}$	0.9940 ± 0.0003	0.9974 ± 0.0002
$\frac{1}{4}$	0.9947 ± 0.0004	0.9974 ± 0.0003
$\frac{1}{2}$	0.9944 ± 0.0004	0.9937 ± 0.0006
1	0.9853 ± 0.0012	0.9700 ± 0.0021
2	0.8988 ± 0.0053	0.9239 ± 0.0050
4	0.7410 ± 0.0072	0.7863 ± 0.0138
8	0.6185 ± 0.0078	0.6467 ± 0.0405

`usps 0v9`

α	Ridge ACC	Isotron ACC
$\frac{1}{8}$	0.9958 ± 0.0001	0.9984 ± 0.0001
$\frac{1}{4}$	0.9958 ± 0.0001	0.9979 ± 0.0001
$\frac{1}{2}$	0.9953 ± 0.0002	0.9966 ± 0.0003
1	0.9871 ± 0.0005	0.9864 ± 0.0007
2	0.9446 ± 0.0012	0.9565 ± 0.0013
4	0.8262 ± 0.0022	0.8768 ± 0.0041
8	0.6872 ± 0.0024	0.8088 ± 0.0291

`mnist 6v7`

Summary thus far

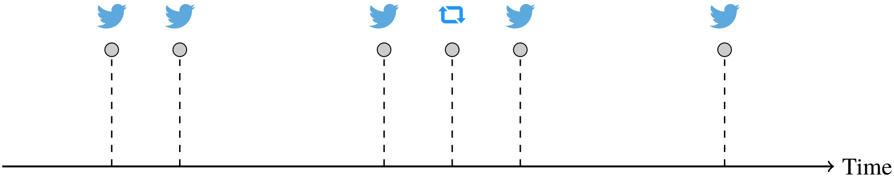


Fitting point processes

Point processes

Model the rate at which events occur in time

- re-tweets in a social network, earthquakes, . . .



Point processes: formally

Suppose $(\mathsf{N}(t))_{t\geq 0}$ counts the # of events in $(0,t]$

Point processes: formally

Suppose $(\mathsf{N}(t))_{t\geq 0}$ counts the # of events in $(0,t]$

In the non-homogeneous Poisson process (NHPP), one posits that the # of events in $(s,t]$ follows

$$\mathsf{N}(t) - \mathsf{N}(s) \sim \text{Poiss}\left(\int_s^t \lambda(u)\,\mathrm{d}u\right)$$

for intensity function $\lambda : \mathbb{R}_+ \to \mathbb{R}_+$

- instantaneous rate at which events occur

NHPP likelihood

Suppose we observe event times $\{t_1, \ldots, t_N\}$, with $T \doteq \max_n t_n$

NHPP likelihood

Suppose we observe event times $\{t_1, \ldots, t_N\}$, with $T \doteq \max_n t_n$

The negative log-likelihood for intensity $\lambda(\cdot;\theta)$ is

$$\mathcal{L}(\theta) \doteq \sum_{n=1}^{N} -\log \lambda(t_n;\theta) + \int_0^T \lambda(u;\theta)\, \mathrm{d}u$$

NHPP likelihood

Suppose we observe event times $\{t_1, \ldots, t_N\}$, with $T \doteq \max_n t_n$

The negative log-likelihood for intensity $\lambda(\cdot;\theta)$ is

$$
\begin{aligned}
\mathcal{L}(\theta) &\doteq \sum_{n=1}^{N} -\log \lambda(t_n;\theta) + \int_0^T \lambda(u;\theta)\,\mathrm{d}u \\
&\propto \frac{1}{N}\sum_{n=1}^{N} -\log \lambda(t_n;\theta) + \frac{T}{N}\cdot \int_0^T \frac{1}{T}\cdot \lambda(u;\theta)\,\mathrm{d}u
\end{aligned}
$$

NHPP likelihood

Suppose we observe event times $\{t_1, \ldots, t_N\}$, with $T \doteq \max_n t_n$

The negative log-likelihood for intensity $\lambda(\cdot; \theta)$ is

$$\begin{aligned}
\mathcal{L}(\theta) &\doteq \sum_{n=1}^{N} -\log \lambda(t_n; \theta) + \int_0^T \lambda(u; \theta)\, \mathrm{d}u \\
&\propto \frac{1}{N} \sum_{n=1}^{N} -\log \lambda(t_n; \theta) + \frac{T}{N} \cdot \int_0^T \frac{1}{T} \cdot \lambda(u; \theta)\, \mathrm{d}u \\
&= \mathop{\mathbb{E}}_{\mathsf{T} \sim \hat{P}} [-\log \lambda(\mathsf{T}; \theta)] + \frac{T}{N} \cdot \mathop{\mathbb{E}}_{\mathsf{T}' \sim Q} \left[\lambda(\mathsf{T}'; \theta)\right]
\end{aligned}$$

where Q is uniform over $[0, T]$

NHPP likelihood

Suppose we observe event times $\{t_1, \ldots, t_N\}$, with $T \doteq \max_n t_n$

The negative log-likelihood for intensity $\lambda(\cdot;\theta)$ is

$$
\begin{aligned}
\mathcal{L}(\theta) &\doteq \sum_{n=1}^{N} -\log \lambda(t_n;\theta) + \int_0^T \lambda(u;\theta)\,\mathrm{d}u \\
&\propto \frac{1}{N}\sum_{n=1}^{N} -\log \lambda(t_n;\theta) + \frac{T}{N} \cdot \int_0^T \frac{1}{T} \cdot \lambda(u;\theta)\,\mathrm{d}u \\
&= \mathop{\mathbb{E}}_{\mathsf{T}\sim\hat{P}} \left[-\log \lambda(\mathsf{T};\theta)\right] + \frac{T}{N} \cdot \mathop{\mathbb{E}}_{\mathsf{T}'\sim Q} \left[\lambda(\mathsf{T}';\theta)\right]
\end{aligned}
$$

where Q is uniform over $[0, T]$

Classification with a uniform background!

NHPPs as binary classification

On an interval $[0,T]$, event times $\{t_1,\ldots,t_N\}$ are iid with density

$$p(t) = \frac{\lambda(t)}{\int_0^T \lambda(u)\,\mathrm{d}u}$$

NHPPs as binary classification

On an interval $[0, T]$, event times $\{t_1, \ldots, t_N\}$ are iid with density

$$p(t) = \frac{\lambda(t)}{\int_0^T \lambda(u)\,\mathrm{d}u}$$

Asymptotically, the likelihood is the classification risk

NHPPs as binary classification

On an interval $[0, T]$, event times $\{t_1, \ldots, t_N\}$ are iid with density

$$p(t) = \frac{\lambda(t)}{\int_0^T \lambda(u)\,\mathrm{d}u}$$

Asymptotically, the likelihood is the classification risk with minimiser

$$\mathop{\mathbb{E}}_{P}[-\log \lambda(\mathsf{T})] + \frac{T}{N} \cdot \mathop{\mathbb{E}}_{Q}\left[\lambda(\mathsf{T}')\right]$$

NHPPs as binary classification

On an interval $[0,T]$, event times $\{t_1,\ldots,t_N\}$ are iid with density

$$\color{blue}{p}(t) = \frac{\lambda(t)}{\int_0^T \lambda(u)\,\mathrm{d}u}$$

Asymptotically, the likelihood is the classification risk with minimiser

$$\underset{\lambda\in\mathbb{R}_+^{\mathcal{X}}}{\operatorname{argmin}} \underset{\color{blue}{P}}{\mathbb{E}}\left[-\log\lambda(\mathsf{T})\right] + \frac{T}{N}\cdot\underset{\color{red}{Q}}{\mathbb{E}}\left[\lambda(\mathsf{T}')\right] = N\cdot\color{blue}{p}$$

NHPPs as binary classification

On an interval $[0, T]$, event times $\{t_1, \ldots, t_N\}$ are iid with density

$$p(t) = \frac{\lambda(t)}{\int_0^T \lambda(u)\,\mathrm{d}u}$$

Asymptotically, the likelihood is the classification risk with minimiser

$$\underset{\lambda \in \mathbb{R}_+^{\mathcal{X}}}{\operatorname{argmin}} \mathop{\mathbb{E}}_{P}[-\log \lambda(\mathsf{T})] + \frac{T}{N} \cdot \mathop{\mathbb{E}}_{Q}\left[\lambda(\mathsf{T}')\right] = (N/T) \cdot p/q$$

NHPPs as binary classification

On an interval $[0, T]$, event times $\{t_1, \ldots, t_N\}$ are iid with density

$$p(t) = \frac{\lambda(t)}{\int_0^T \lambda(u)\,\mathrm{d}u}$$

Asymptotically, the likelihood is the classification risk with minimiser

$$\underset{\lambda \in \mathbb{R}_+^{\mathcal{X}}}{\operatorname{argmin}} \underset{P}{\mathbb{E}}[-\log \lambda(\mathsf{T})] + \frac{T}{N} \cdot \underset{Q}{\mathbb{E}}\left[\lambda(\mathsf{T}')\right] = (N/T) \cdot p/q$$

Weighted density ratio estimation!

Generalised likelihood?

For scorer $s \colon \mathbb{R}_+ \to \mathbb{R}$, consider

$$\begin{aligned} & \min_{s \in \mathcal{S}} \mathop{\mathbb{E}}_{\hat{P}} \left[\ell(+1, s(\mathsf{T}))\right] + \frac{T}{N} \cdot \mathop{\mathbb{E}}_{Q} \left[\ell(-1, s(\mathsf{T}'))\right] \\ &= \min_{s \in \mathcal{S}} \sum_{n=1}^{N} \ell(+1, s(t_n)) + \int_0^T \ell(-1, s(t))\, \mathrm{d}t \end{aligned}$$

for strictly proper composite ℓ

Generalised likelihood?

For scorer $s \colon \mathbb{R}_+ \to \mathbb{R}$, consider

$$\min_{s \in \mathcal{S}} \mathbb{E}_{\hat{P}} \left[\ell(+1, s(\mathsf{T}))\right] + \frac{T}{N} \cdot \mathbb{E}_{Q} \left[\ell(-1, s(\mathsf{T}'))\right]$$
$$= \min_{s \in \mathcal{S}} \sum_{n=1}^{N} \ell(+1, s(t_n)) + \int_0^T \ell(-1, s(t))\, \mathrm{d}t$$

for strictly proper composite ℓ

We retain the optimal solution by picking

$$\lambda(t) = \frac{\Psi^{-1}(s(t))}{1 - \Psi^{-1}(s(t))}$$

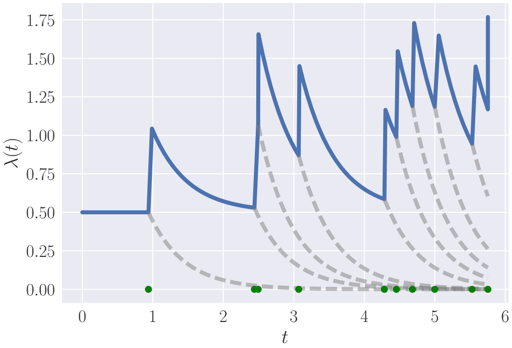
- optimal $s = \Psi(\eta)$, $\frac{\eta}{1-\eta} \propto p$

Application: Hawkes processes

The self-exciting Hawkes process assumes, for link $F(\cdot)$,

$$\lambda\left(t;\{t_n\}_{n=1}^{N}\right) = F\left(\mu + \alpha \cdot \sum_{t_n<t} e^{-\delta\cdot(t-t_n)}\right)$$

- occurrence of one event triggers subsequent events



Generalised Hawkes likelihood?

In terms of a scorer, the Hawkes intensity is

$$\begin{aligned} \lambda\left(t; \{t_n\}_{n=1}^{N}\right) &= F\left(s(t)\right) \\ s(t) &= \mu + \alpha \cdot \Phi(t) \\ \Phi(t) &\doteq \sum_{t_n < t} e^{-\delta \cdot (t - t_n)} \end{aligned}$$

Menon and Lee. Beyond the likelihood: new loss functions for (non-)linear Hawkes processes. In preparation.

Generalised Hawkes likelihood?

In terms of a scorer, the Hawkes intensity is

$$\begin{aligned} \lambda\left(t; \{t_n\}_{n=1}^{N}\right) &= F\left(s(t)\right) \\ s(t) &= \mu + \alpha \cdot \Phi(t) \\ \Phi(t) &\doteq \sum_{t_n < t} e^{-\delta \cdot (t - t_n)} \end{aligned}$$

Can minimise a proper loss with this $s(\cdot)$ and Φ, and set

$$\lambda(t) = \frac{\Psi^{-1}(s(t))}{1 - \Psi^{-1}(s(t))}$$

Menon and Lee. Beyond the likelihood: new loss functions for (non-)linear Hawkes processes. In preparation.

Generalised Hawkes likelihood?

In terms of a scorer, the Hawkes intensity is

$$\begin{aligned} \lambda\left(t; \{t_n\}_{n=1}^{N}\right) &= F\left(s(t)\right) \\ s(t) &= \mu + \alpha \cdot \Phi(t) \\ \Phi(t) &\doteq \sum_{t_n < t} e^{-\delta \cdot (t - t_n)} \end{aligned}$$

Can minimise a proper loss with this $s(\cdot)$ and Φ, and set

$$\lambda(t) = \frac{\Psi^{-1}(s(t))}{1 - \Psi^{-1}(s(t))} = F(s(t))$$

if we choose

$$\Psi^{-1}(v) = \frac{F(v)}{1 + F(v)}$$

Menon and Lee. Beyond the likelihood: new loss functions for (non-)linear Hawkes processes. In preparation.

Hawkes process with linear $F(\cdot)$

For $F(z) = z$, we may explore losses with $\Psi(u) = \frac{u}{1-u}$

- losses that directly seek density ratios

Hawkes process with linear $F(\cdot)$

For $F(z) = z$, we may explore losses with $\Psi(u) = \frac{u}{1-u}$

- losses that directly seek density ratios

One appealing candidate (Kanamori et al., 2009):

$$\ell(+1, v) = -v \qquad \ell(-1, v) = \frac{1}{2}v^2$$

- c.f. (Reynaud-Bouret 2014, Bacry et al., 2015)

Hawkes process with linear $F(\cdot)$

For $F(z) = z$, we may explore losses with $\Psi(u) = \frac{u}{1-u}$

- losses that directly seek density ratios

One appealing candidate (Kanamori et al., 2009):

$$\ell(+1, v) = -v \qquad \ell(-1, v) = \frac{1}{2}v^2$$

- c.f. (Reynaud-Bouret 2014, Bacry et al., 2015)

Potential closed-form solution

$$\theta^* = \frac{N}{T} \cdot \left(\mathbb{E}_{Q}\left[\Phi(\mathsf{T}')\Phi(\mathsf{T}')^T\right] \right)^{-1} \mathbb{E}_{\hat{P}}[\Phi(\mathsf{T})]$$

when this quantity is non-negative

Hawkes process with exponential $F(\cdot)$

For $F(z) = e^z$, we may explore losses with $\Psi(u) = \log \frac{u}{1-u}$

Hawkes process with exponential $F(\cdot)$

For $F(z) = e^z$, we may explore losses with $\Psi(u) = \log \frac{u}{1-u}$

One appealing candidate is familiar logistic loss

- nonlinear Hawkes with logistic regression!

Hawkes process with exponential $F(\cdot)$

For $F(z) = e^z$, we may explore losses with $\Psi(u) = \log \frac{u}{1-u}$

One appealing candidate is familiar logistic loss

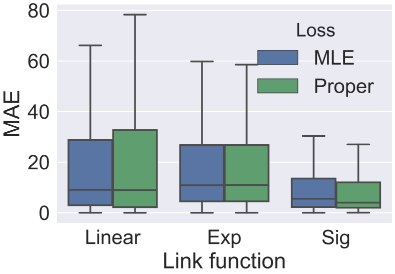
- nonlinear Hawkes with logistic regression!

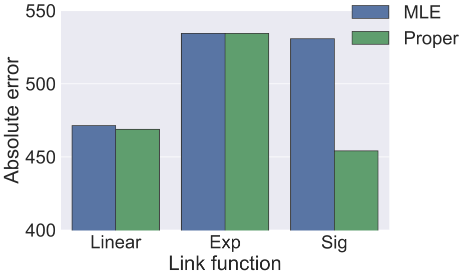
By weighting the negative class, this is actually equivalent to MLE

- follows from (Fithian & Hastie, 2013)

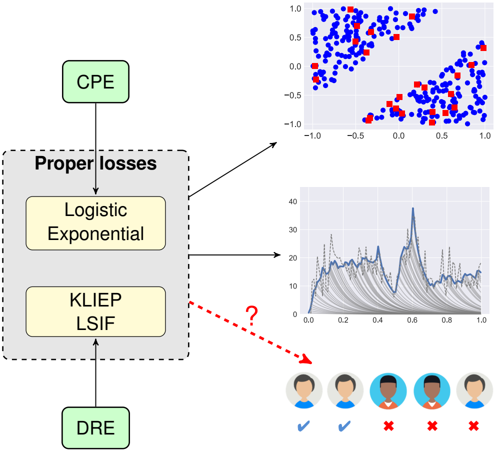
Illustration: fitting with proper losses

Prediction of # events on `lastfm` and `bitcoin` datasets





Summary thus far

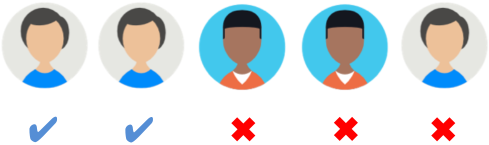


Fairness-aware classification

Fairness-aware classification

Learn a classifier achieving two goals:

- accurately predict a target label
- don't discriminate on some sensitive feature



Fairness-aware classification: formally

We seek a classifier $f\colon \mathcal{X} \to \{\pm 1\}$, with induced predictions $\hat{\mathsf{Y}}$

Fairness-aware classification: formally

We seek a classifier $f : \mathcal{X} \to \{\pm 1\}$, with induced predictions $\hat{\mathsf{Y}}$

f should predict well target variable Y

- e.g. attain low balanced error,

$$\mathrm{BER}(f) \doteq \frac{1}{2} \cdot \left(\mathbb{P}(\hat{\mathsf{Y}} = +1 \mid \mathsf{Y} = -1) + \mathbb{P}(\hat{\mathsf{Y}} = -1 \mid \mathsf{Y} = +1) \right)$$

Fairness-aware classification: formally

We seek a classifier $f \colon \mathcal{X} \to \{\pm 1\}$, with induced predictions $\hat{\mathsf{Y}}$

f should predict well target variable Y

- e.g. attain low balanced error,

$$\mathrm{BER}(f) \doteq \frac{1}{2} \cdot \left(\mathbb{P}(\hat{\mathsf{Y}} = +1 \mid \mathsf{Y} = -1) + \mathbb{P}(\hat{\mathsf{Y}} = -1 \mid \mathsf{Y} = +1) \right)$$

f should predict poorly sensitive variable $\overline{\mathsf{Y}}$

- e.g. attain high balanced error,

$$\overline{\mathrm{BER}}(f) \doteq \frac{1}{2} \cdot \left(\mathbb{P}(\hat{\mathsf{Y}} = +1 \mid \overline{\mathsf{Y}} = -1) + \mathbb{P}(\hat{\mathsf{Y}} = -1 \mid \overline{\mathsf{Y}} = +1) \right)$$

Fairness-aware objective

We seek a solution to

$$\min_f \mathrm{BER}(f) - \lambda \cdot \overline{\mathrm{BER}}(f)$$

Fairness-aware objective

We seek a solution to

$$\min_f \mathrm{BER}(f) - \lambda \cdot \overline{\mathrm{BER}}(f) = \min_s \mathop{\mathbb{E}}_{P} [\![s(\mathsf{X}) < 0]\!] + \mathop{\mathbb{E}}_{Q} [\![s(\mathsf{X}') > 0]\!]$$

$$- \lambda \cdot \left(\mathop{\mathbb{E}}_{\overline{P}} [\![s(\mathsf{X}) < 0]\!] + \mathop{\mathbb{E}}_{\overline{Q}} [\![s(\mathsf{X}') > 0]\!] \right)$$

Fairness-aware objective

We seek a solution to

$$\min_f \mathrm{BER}(f) - \lambda \cdot \overline{\mathrm{BER}}(f) = \min_s \mathop{\mathbb{E}}_{P} [\![s(\mathsf{X}) < 0]\!] + \mathop{\mathbb{E}}_{Q} [\![s(\mathsf{X}') > 0]\!]$$
$$- \lambda \cdot \left(\mathop{\mathbb{E}}_{\overline{P}} [\![s(\mathsf{X}) < 0]\!] + \mathop{\mathbb{E}}_{\overline{Q}} [\![s(\mathsf{X}') > 0]\!] \right)$$

Natural to consider surrogate risk

$$\min_s \mathrm{BER}_\ell(s) - \lambda \cdot \overline{\mathrm{BER}}_\ell(s)$$

Fairness-aware objective

We seek a solution to

$$\min_{f} \mathrm{BER}(f) - \lambda \cdot \overline{\mathrm{BER}}(f) = \min_{s} \mathop{\mathbb{E}}_{P} [\![s(\mathsf{X}) < 0]\!] + \mathop{\mathbb{E}}_{Q} [\![s(\mathsf{X}') > 0]\!]$$
$$- \lambda \cdot \left(\mathop{\mathbb{E}}_{\overline{P}} [\![s(\mathsf{X}) < 0]\!] + \mathop{\mathbb{E}}_{\overline{Q}} [\![s(\mathsf{X}') > 0]\!] \right)$$

Natural to consider surrogate risk

$$\min_{s} \mathrm{BER}_{\ell}(s) - \lambda \cdot \overline{\mathrm{BER}}_{\ell}(s) = \min_{s} \mathop{\mathbb{E}}_{P} \ell(+1, s(\mathsf{X})) + \mathop{\mathbb{E}}_{Q} \ell(-1, s(\mathsf{X}'))$$
$$- \lambda \cdot \left(\mathop{\mathbb{E}}_{\overline{P}} \ell(+1, s(\mathsf{X})) + \mathop{\mathbb{E}}_{\overline{Q}} \ell(-1, s(\mathsf{X}')) \right)$$

Fairness-aware objective

We seek a solution to

$$\min_f \mathrm{BER}(f) - \lambda \cdot \overline{\mathrm{BER}}(f) = \min_s \mathop{\mathbb{E}}_{P} [\![s(\mathsf{X}) < 0]\!] + \mathop{\mathbb{E}}_{Q} [\![s(\mathsf{X}') > 0]\!] - \lambda \cdot \left(\mathop{\mathbb{E}}_{\overline{P}} [\![s(\mathsf{X}) < 0]\!] + \mathop{\mathbb{E}}_{\overline{Q}} [\![s(\mathsf{X}') > 0]\!] \right)$$

Natural to consider surrogate risk

$$\min_s \mathrm{BER}_\ell(s) - \lambda \cdot \overline{\mathrm{BER}}_\ell(s) = \min_s \mathop{\mathbb{E}}_{P} \ell(+1, s(\mathsf{X})) + \mathop{\mathbb{E}}_{Q} \ell(-1, s(\mathsf{X}')) - \lambda \cdot \left(\mathop{\mathbb{E}}_{\overline{P}} \ell(+1, s(\mathsf{X})) + \mathop{\mathbb{E}}_{\overline{Q}} \ell(-1, s(\mathsf{X}')) \right)$$

but in general this will be non-convex

CPE approach?

Alternately, let us consider the Bayes-optimal solutions

$$f^* \in \underset{f \colon \mathcal{X} \to \{\pm 1\}}{\operatorname{argmin}} \mathrm{BER}(f) - \lambda \cdot \overline{\mathrm{BER}}(f)$$

Menon and Williamson. The cost of fairness in binary classification. `https://arxiv.org/abs/1705.09055`

CPE approach?

Alternately, let us consider the Bayes-optimal solutions

$$f^* \in \operatorname*{argmin}_{f\colon \mathcal{X} \to \{\pm 1\}} \mathrm{BER}(f) - \lambda \cdot \overline{\mathrm{BER}}(f)$$

Easy to show that

$$f^*(x) = [\![\eta(x) - \pi > \lambda \cdot (\overline{\eta}(x) - \overline{\pi})]\!]$$

$$\eta(x) \doteq \mathbb{P}(\mathsf{Y} = +1 \mid \mathsf{X} = x)$$
$$\overline{\pi} \doteq \mathbb{P}(\overline{\mathsf{Y}} = +1)$$

Menon and Williamson. The cost of fairness in binary classification. `https://arxiv.org/abs/1705.09055`

CPE approach?

Alternately, let us consider the Bayes-optimal solutions

$$f^* \in \underset{f\colon \mathcal{X} \to \{\pm 1\}}{\operatorname{argmin}} \operatorname{BER}(f) - \lambda \cdot \overline{\operatorname{BER}}(f)$$

Easy to show that

$$f^*(x) = [\![\eta(x) - \pi > \lambda \cdot (\bar{\eta}(x) - \bar{\pi})]\!]$$

$$\eta(x) \doteq \mathbb{P}(\mathsf{Y} = +1 \mid \mathsf{X} = x) \qquad \bar{\eta}(x) \doteq \mathbb{P}(\bar{\mathsf{Y}} = +1 \mid \mathsf{X} = x)$$

$$\bar{\pi} \doteq \mathbb{P}(\bar{\mathsf{Y}} = +1) \qquad \pi \doteq \mathbb{P}(\mathsf{Y} = +1)$$

Menon and Williamson. The cost of fairness in binary classification. https://arxiv.org/abs/1705.09055

CPE approach?

Alternately, let us consider the Bayes-optimal solutions

$$f^* \in \underset{f \colon \mathcal{X} \to \{\pm 1\}}{\operatorname{argmin}} \operatorname{BER}(f) - \lambda \cdot \overline{\operatorname{BER}}(f)$$

Easy to show that

$$f^*(x) = [\![\eta(x) - \pi > \lambda \cdot (\overline{\eta}(x) - \overline{\pi})]\!]$$

$$\eta(x) \doteq \mathbb{P}(\mathsf{Y} = +1 \mid \mathsf{X} = x) \qquad \overline{\eta}(x) \doteq \mathbb{P}(\overline{\mathsf{Y}} = +1 \mid \mathsf{X} = x)$$

$$\overline{\pi} \doteq \mathbb{P}(\overline{\mathsf{Y}} = +1) \qquad \pi \doteq \mathbb{P}(\mathsf{Y} = +1)$$

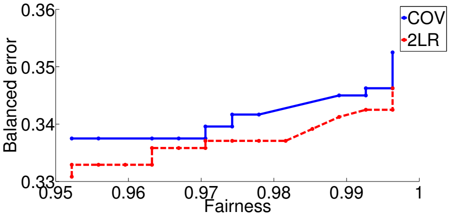
Just requires CPE on the target and sensitive features!

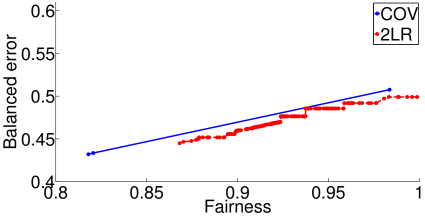
- tuning of λ does not require re-training
- also useful to study feature learning (McNamara et al., 2017)

Menon and Williamson. The cost of fairness in binary classification. `https://arxiv.org/abs/1705.09055`

Illustration of CPE approach

Competitive performance with bespoke optimisation (`COV`) on UCI `adult` and synthetic Gaussian datasets



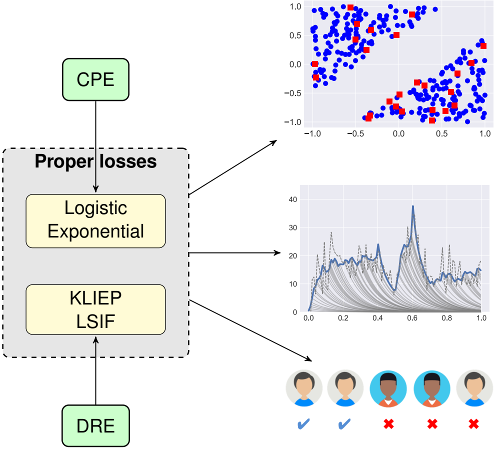


Conclusion

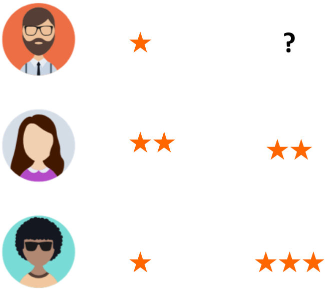
Talk summary

A formal link between DRE and CPE

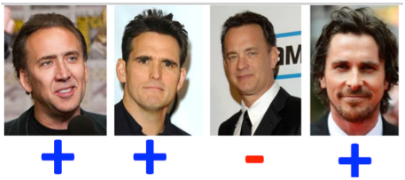
CPE approach to three distinct learning problems



For another day



Recommender systems



Ranking

Collaborators

Brendan van Rooyen
ANU

Bob Williamson
ANU

Cheng Soon Ong
Data61/ANU

Richard Nock
Data61/ANU

Nagarajan Natarajan
MSR Bangalore

Giorgio Patrini
UvA-Bosch DELTA

Young Lee
Data61/ANU

Lizhen Qu
Data61/ANU

Thanks!

Further reading

Linking losses for density ratio and class-probability estimation. Aditya Krishna Menon and Cheng Soon Ong. ICML 2016.

A scaled Bregman theorem with applications. Richard Nock, Aditya Krishna Menon and Cheng Soon Ong. NIPS 2016.

Learning from corrupted binary labels via class-probability estimation. Aditya Krishna Menon, Brendan van Rooyen, Cheng Soon Ong and Robert C. Williamson. ICML 2015.

Learning with symmetric label noise: the importance of being unhinged. Brendan van Rooyen, Aditya Krishna Menon and Robert C. Williamson. NIPS 2015.

Learning from binary labels with instance-dependent corruption. Aditya Krishna Menon, Brendan van Rooyen and Nagarajan Natarajan. https://arxiv.org/abs/1605.00751

Making deep neural networks robust to label noise: a loss correction approach. Giorgio Patrini, Alessandro Rozza, Aditya Krishna Menon, Richard Nock, Lizhen Qu. CVPR 2017.

Beyond the likelihood: new loss functions for (non-)linear Hawkes processes. Aditya Krishna Menon and Young Lee. In preparation.

The cost of fairness in binary classification. Aditya Krishna Menon and Robert C. Williamson. https://arxiv.org/abs/1705.09055